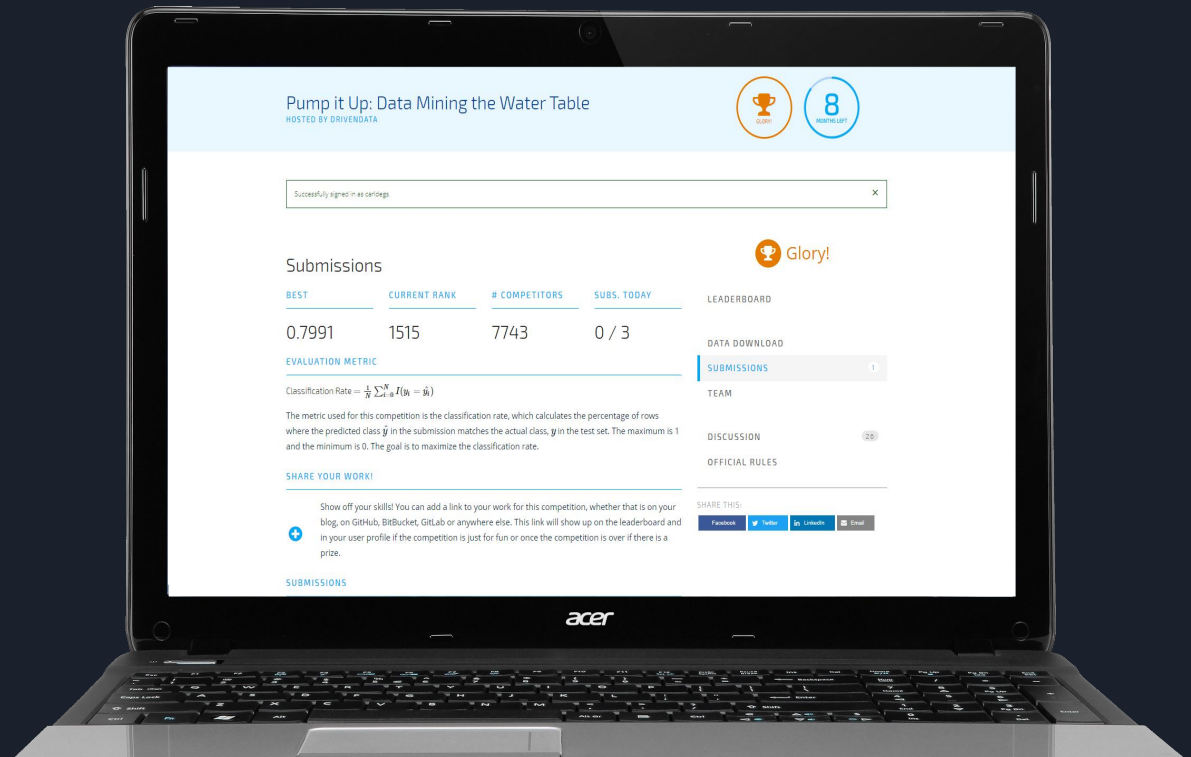
Pump it Up

Data Mining the Water Table

About the Contest



Hosted at
drivendata.org

Dataset provided by
Taarifa

Number of Teams
7743

Objective

Predict the operating condition of a waterpoint, based on the data from Taarifa and the Tanzanian Ministry of Water.

Importance

A smart understanding of which waterpoints will fail can improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania.

Data

59,400

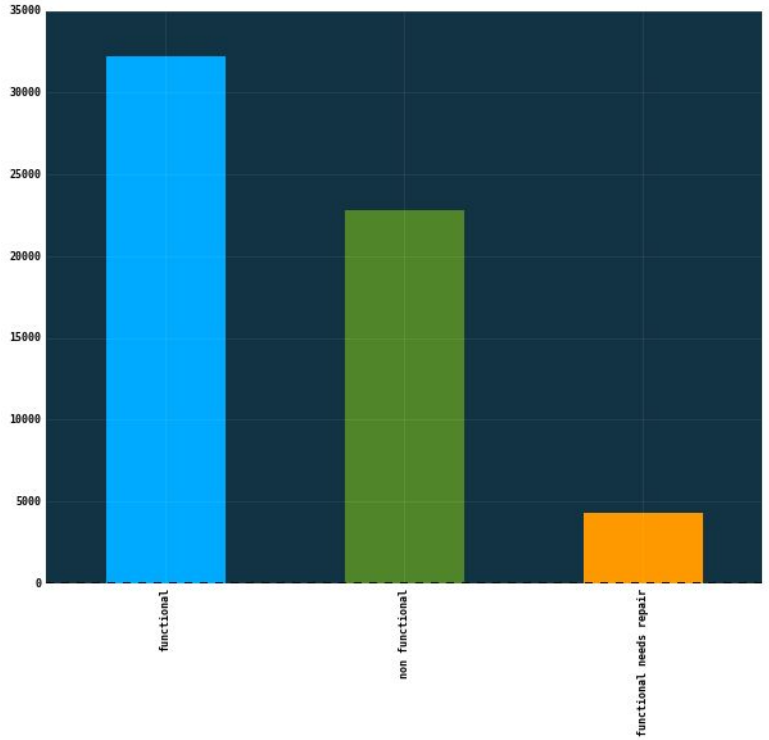
Datapoints

40

Features

Functional, Non-functional, Functional but needs repair

Labels



Preprocessing

Remove all redundant, numerous, unusable or unnecessary data

Unique data
Duplicate data
Singular value
Numerous data

amount_tsh
date_recorded
funder*
gps_height
installer*
longitude
latitude
wpt_name
num_private
basin
subvillage
region
region_code
district_code
lga*
ward*
population
public_meeting
Recorded_by
scheme_management
scheme_name
permit
construction_year
extraction_type
extraction_type_group
extraction_type_class
management
management_group
payment
payment_type
water_quality
quality_group
quantity
quantity_group
source
source_type
waterpoint_type
waterpoint_type_group

Preprocessing

Divide the data into **ordinal** and **nominal**.

amount_tsh
date_recorded
gps_height
longitude
latitude
basin
region_code
district_code
population
public_meeting

scheme_management
permit
construction_year
extraction_type
extraction_type_class
management
management_group
payment_type
water_quality
quality_group
quantity
source
waterpoint_type

Preprocessing

Ordinal data are scaled [0,1]

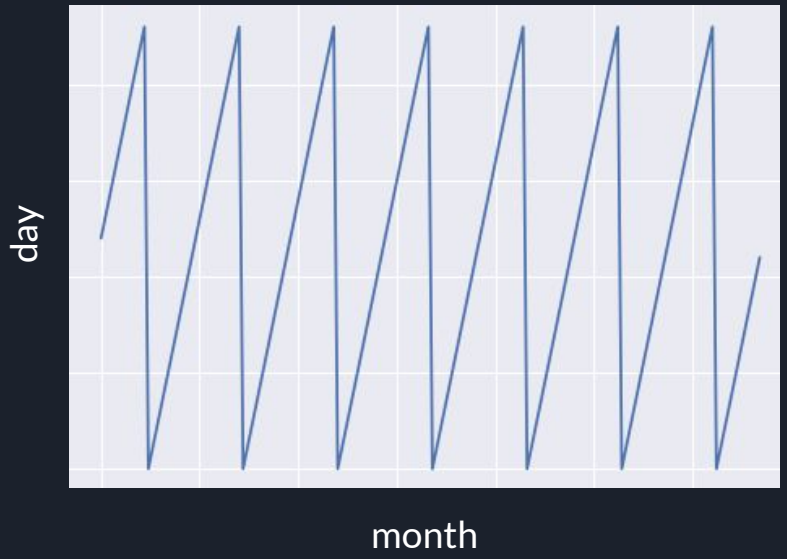
Nominal data are one-hot encoded

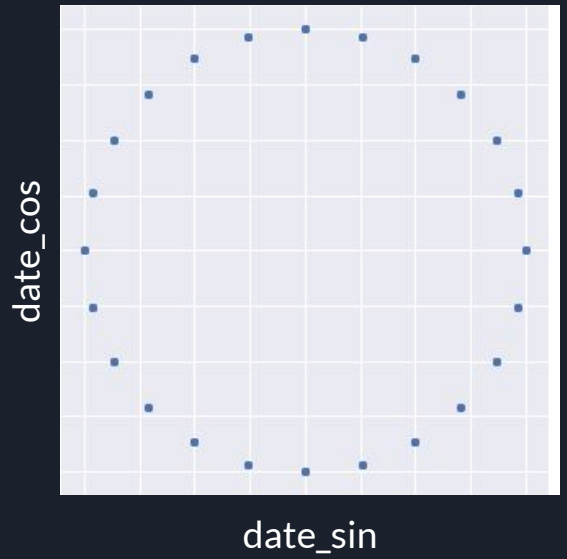
amount_tsh
date_recorded
gps_height
longitude
latitude
population

basin
region_code
district_code
scheme_management
permit
public_meeting
construction_year
extraction_type
extraction_type_class
management
management_group
payment_type
water_quality
quality_group
quantity
source
waterpoint_type

Preprocessing
Other methods tested

Treat dates (e.g. **date_recorded**) as cyclical data.





Preprocessing
Other methods tested

Use a feature hasher on numerous features with no fixed set of values (e.g. **funder** and **installer**)

Model and Implementation

3-Layer MLP (1024 neurons) with Activation of ReLu and Dropout of 0.4 after each layer

Activation Layers: ReLU (empirical)

Dropout: 0.4 (empirical)

Adam optimizer (empirical)

Method and Rationale

Present the data in a meaningful way such that the Neural Network can be able to learn from it.

Then let the Neural Network learn

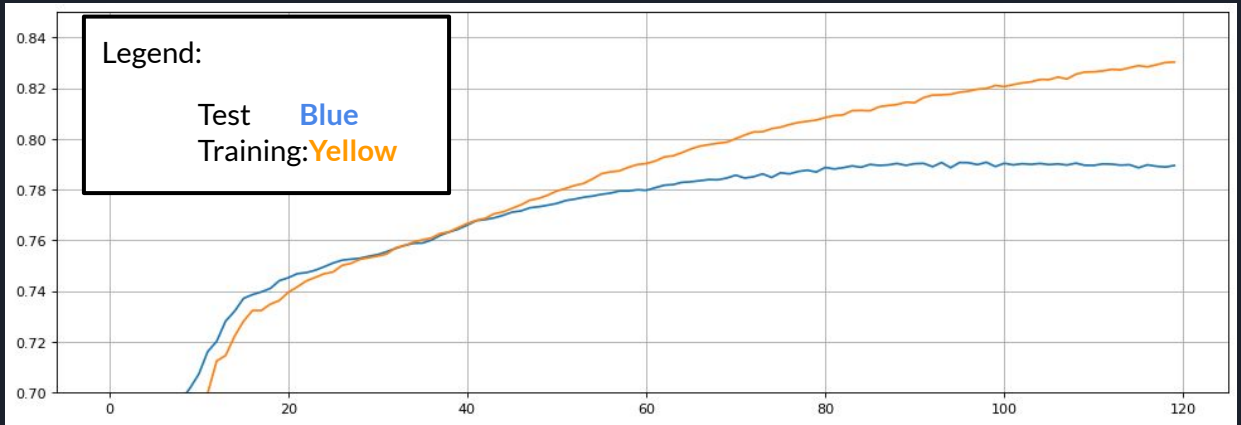
I.e. Transforming the nominal data into one-hot encoded data allows the NN to differentiate between different classes such as waterpoints managed by government and those managed by others. Maintaining numerical data allows magnitude to play a role (e.g. amount_tsh)

Some data seemed faulty and we tried “cleaning” them by replacing the faulty values with the mean

Approach
K-fold validation

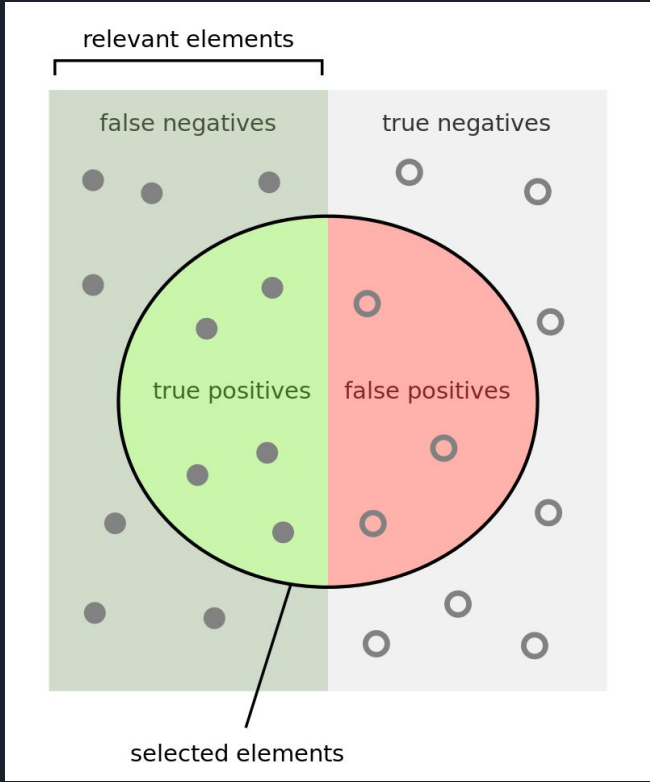
K-fold validation to be able to roughly predict the score by DrivenData.Org

By observing the K-fold validation accuracy, the epochs for the final training on the entire data set was set.



Precision vs. Recall

https://en.wikipedia.org/wiki/Precision_and_recall



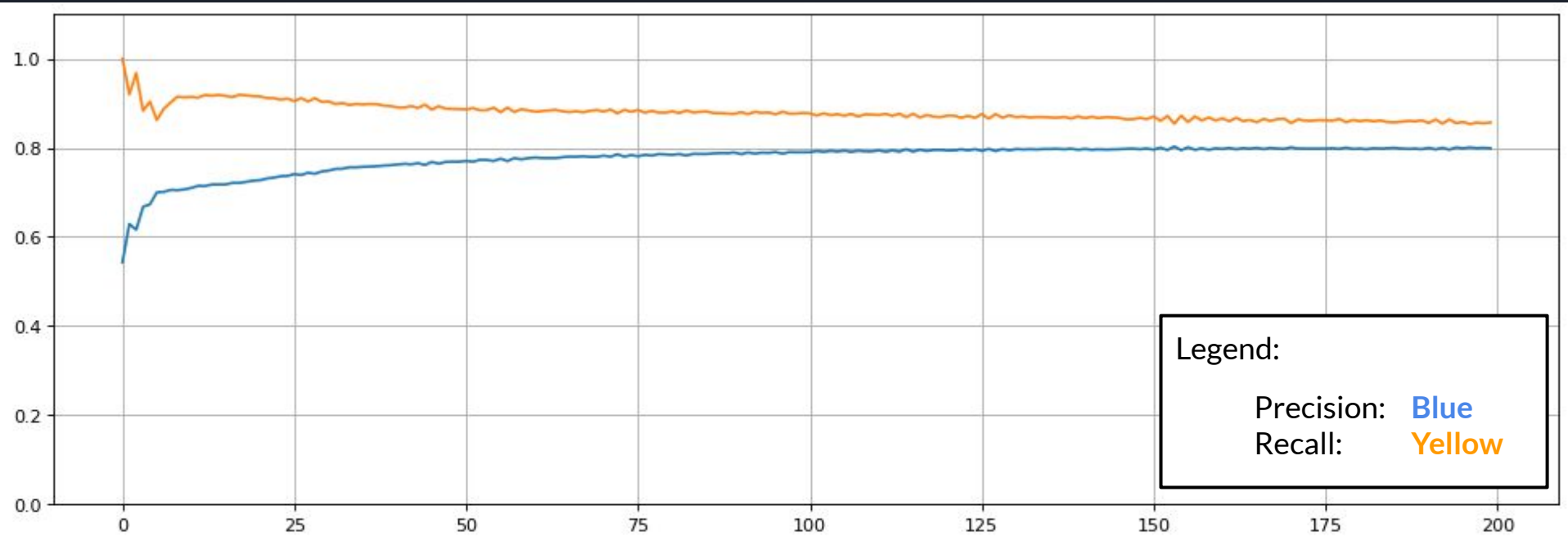
How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}}$$

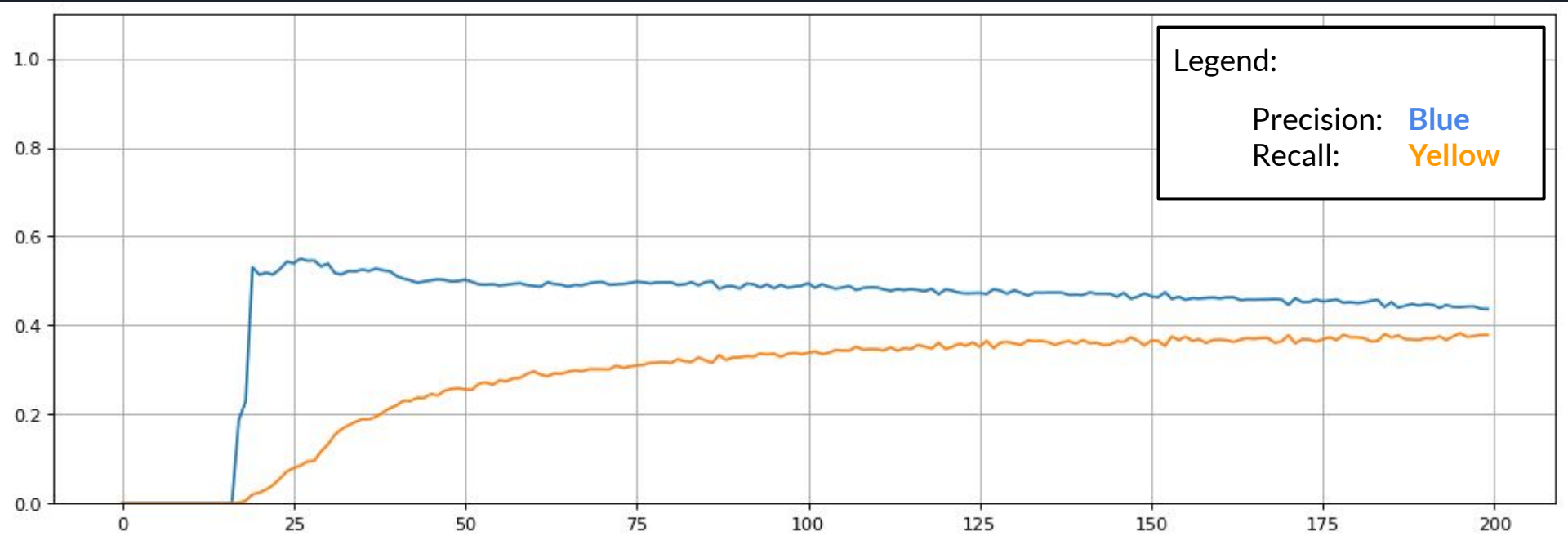
How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

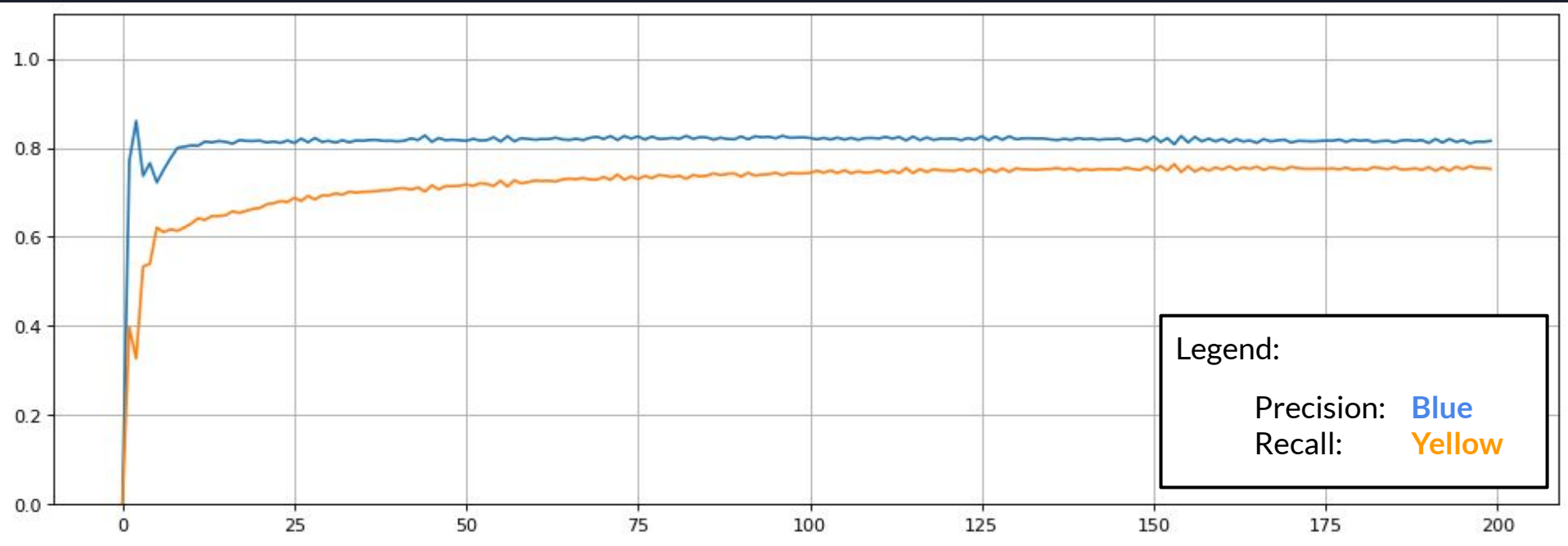
Precision and Recall curve (Functional)



Precision and Recall curve (Functional Needs Repair)



Precision and Recall curve (Non Functional)



Analysis and Discussion

The Recall and Precision plots seem to indicate that to obtain further improvements. Some form of data augmentation would be necessary.

The leaderboard’s highest score is 82.86%

Possibly because of good “functional needs repair” classification

Analysis and Discussion

Attempts at filtering the data columns generally negatively affected performance.

Human filtering only seemed to work for

- Duplicate Columns
- Columns with unique values for all points
- Columns with one value for all points

Analysis and Discussion

The Recall and Precision plots seem to indicate that to obtain further improvements. Some form of data augmentation would be necessary.

The performance on the two classes with plenty of data is satisfactory (functional & non functional. But much is left to be desired for the class with little data (functional needs repair)

The leaderboards indicate that 83% seems to be State of the Art for this competition. This 3% could very well be due to a significantly better performance classifying “functional but needs repair”

Results

79.91%

Classification Rate

Top 20% of all the submissions

Rank when this was submitted